Songyi AHN, Nakyung LEE, Nishat PARVEZ

Professor Joseph NESE

EDLD 651

**Final Project Outline**

**I. Introduction**

The increasing use of Telehealth services has redefined healthcare delivery, offering convenient and accessible options for a wide range of health needs. However, understanding who benefits from these services and the factors influencing their usage remains essential for enhancing equitable healthcare access. This study investigates demographic, socio-economic, and behavioral characteristics that may shape individuals' use of Telehealth services, with a focus on identifying key differences between those who use Telehealth and those who do not.

Three research questions guide this analysis. First, the study examines how demographic and socio-economic factors—including age, gender, ethnicity, education, and income—differ between Telehealth users and non-users. This analysis can reveal patterns of healthcare access, highlighting potential disparities in digital health engagement. Second, the study explores social media usage patterns and engagement levels with health information on social media, comparing these between individuals based on their Telehealth usage. Since social media is a major source of health information, understanding how social media engagement levels are associated with Telehealth adoption can inform digital health outreach strategies. Lastly, the study assesses how cancer-related risk perceptions vary between Telehealth users and non-users. As Telehealth can play a critical role in preventive care and health education, analyzing cancer risk perceptions can shed light on how effectively it supports awareness and prevention.

This study aims to uncover demographic and behavioral factors associated with Telehealth use using a dataset from the National Cancer Institute, which captures comprehensive health information-seeking behaviors and health technology use across the United States. The findings have potential implications for public health interventions focused on enhancing digital health equity and improving health outcomes through targeted engagement.

**II. Research Questions**

*RQ1:* What is the distribution of demographic/socio-economic categories, including age, gender, ethnicity, education, and income level, among individuals who do and do not use Telehealth services?

*RQ2:* How do social media usage frequency and health information engagement levels on social media (low, medium, high) differ between individuals who use and do not use Telehealth services?

*RQ3:* How does the distribution of cancer-related risk perception groups (low, medium, high) compare between individuals who use and do not use Telehealth services?

**III. Data Description**

This dataset captures detailed information on health information-seeking behavior, digital health technology use, cancer awareness, and preventive health practices across the United States. Collected by the National Cancer Institute, this dataset includes a diverse range of variables organized into several key categories:

* **Demographics:** This section includes respondents’ age, gender, race/ethnicity, education, income, employment, and location, allowing analysis across various subgroups to reveal disparities in health information access.
* **Health Information Seeking:** Capturing where and how often individuals look for health information, this category includes sources such as healthcare providers, the internet, social media, and personal networks, along with respondents’ trust levels for each source.
* **Digital Health Technology:** This section assesses respondents’ usage of health apps, wearable devices, and online health portals, illustrating how technology is used to manage personal health and access medical information.
* **Cancer Awareness:** With variables measuring knowledge of cancer risks, prevention, and screening behaviors, this section also includes questions on family cancer history and perceived susceptibility to cancer.
* **Health Literacy:** This measures respondents’ self-assessed ability to understand and navigate health information, critical for making informed health decisions.
* **COVID-19 Impact:** This section examines changes in health behaviors and attitudes towards health messaging due to the pandemic, particularly in preventive practices and public health trust.
* **Preventive Health Practices:** This category captures self-reported health behaviors such as smoking, alcohol use, exercise, diet, and routine check-ups, providing a broad view of respondents' health maintenance activities.

The dataset is organized in a structured format with metadata explaining each variable’s coding and response options. It enables analyses across demographic segments and is instrumental for public health research, specifically in areas like health literacy, digital health trends, and the impact of media on health decisions.

**IV. Preparatory Work**

Some data preparation steps would be:

(a) Data Cleaning

Identify and decide how to handle missing values, depending on the proportion and importance of each variable.

Standardize formats and numeric scales of data for consistency.

Check for and address any outliers that may affect the analysis.

(b) Encoding Categorical Variables

Convert categorical variables such as gender, ethnicity, and education into numerical codes or factors suitable for analysis.

(c) Data Transformation

Combine, regenerate, or transform necessary data into new variables, which can be our interested variables for the research questions, such as health information engagement scores (health information seeking + health literacy) or cancer-related behaviors (cancer risk perceptions + cancer preventive activities).

(d) Exploratory Data Analysis

Conduct initial visualizations to understand variable distributions (histograms, box plots).

Calculate descriptive statistics (mean, median, mode) for key variables like age, income, and the extent of health information seeking.

Examine some correlation analysis among interested variables to identify if there are no unexpected problems for data analysis to address our research questions.

**V. Meeting Final Project Requirements**

To satisfy the final project requirements:

(a) Reproducible R Markdown Document with References:

We will use R Markdown to create a reproducible document, integrating data analysis, code, and references to relevant literature.

(b) GitHub Repository:

The project will be managed on GitHub, with individual contributions documented through clear commit histories to ensure transparency in each team member’s work.

(c) Data Tidying:

The **pivot\_longer()** function can be used to reshape Health Information Seeking variables into a long format, with a single column representing different types of health information engagement on social media. Specifically, columns, including the frequency of watching health-related videos, interacting with people with similar health issues, and sharing personal or general health information, can be combined into an “engagement type” column, with their respective frequency level in a “frequency” column. After that, we can use the **mutate()** function to calculate the overall engagement score by summing the frequencies for each individual, then categorize this score as “low,” “medium,” and “high.” After creating these categories with mutate(), we can apply the **group\_by()** function with Telehealth usage and the new categories to examine how Telehealth usage differs across the three social media health information engagement level groups.

We will use the **select()** function to isolate all relevant columns from the dataset, including age, gender, ethnicity, income, education level, social media frequency, health information engagement levels on social media, and cancer-related risk perceptions. The **filter()** function will be applied, for example, to focus on responses that match any value representing “Yes” for Telehealth usage, allowing us to analyze these variables specifically among Telehealth users.

The **pivot\_wider()** function can be used to reshape an existing column representing cancer risk perception into three separate columns for “low,” “medium,” and “high” groups by categorizing responses into each range. The mutate() function will be used to classify the risk perception scores into the three categories, and after categorization, pivot\_wider() reshapes the categories into separate columns. This allows us to create a table showing how counts vary across these three risk perception categories when Telehealth usage is “Yes.”

The **summarize()** will be applied to calculate counts or percentages for each demographic factor among Telehealth users and non-users. We can also obtain the median income level for Telehealth users. Furthermore, average values for social media usage frequencies, social media health information engagement levels, cancer risk perceptions can be calculated with the summarize() function.

(d) Exploratory Data Visualizations:

For RQ1, bar charts could illustrate the distribution of demographic categories between Telehealth users and non-users. For RQ2, the distribution of social media usage frequency and social media health information engagement level (low, medium, high) between Telehealth users and non-users will be displayed with a single bar chart and grouped or stacked bar chart, respectively. Similarly, for RQ3, grouped or stacked bar charts can represent the cancer risk perception categories (low, medium, high) for Telehealth users and non-users. Additionally, a pie chart can visualize the proportion of each risk level within Telehealth users group.

(e) Summary Statistics:

Summary statistics can include (1) the number or proportion of individuals by demographic categories such as age group, gender, ethnicity, education, and income level among Telehealth users and non-users; (2) median values for income and age; (3) mean, median, and standard deviation values for social media usage frequency, health information social media engagement level, and cancer risk perception level.